Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(11) 3294–3307


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280220924680
journals.sagepub.com/home/smm
SAGE

# Integrating latent classes in the Bayesian shared parameter joint model of longitudinal and survival outcomes

**Eleni-Rosalina Andrinopoulou[1], Kazem Nasserinejad[2], Rhonda Szczesniak[3,4] and Dimitris Rizopoulos[1]**

**Abstract**

Cystic fibrosis is a chronic lung disease requiring frequent lung-function monitoring to track acute respiratory events (pulmonary exacerbations). The association between lung-function trajectory and time-to-first exacerbation can be characterized using joint longitudinal-survival modeling. Joint models specified through the shared parameter framework quantify the strength of association between such outcomes but do not incorporate latent sub-populations reflective of heterogeneous disease progression. Conversely, latent class joint models explicitly postulate the existence of sub-populations but do not directly quantify the strength of association. Furthermore, choosing the optimal number of classes using established metrics like deviance information criterion is computationally intensive in complex models. To overcome these limitations, we integrate latent classes in the shared parameter joint model through a fully Bayesian approach. To choose the optimal number of classes, we construct a mixture model assuming more latent classes than present in the data, thereby asymptotically "emptying" superfluous latent classes, provided the Dirichlet prior on class proportions is sufficiently uninformative. Model properties are evaluated in simulation studies. Application to data from the US Cystic Fibrosis Registry supports the existence of three sub-populations corresponding to lung-function trajectories with high initial forced expiratory volume in 1 s ($FEV_1$), rapid $FEV_1$ decline, and low but steady $FEV_1$ progression. The association between $FEV_1$ and hazard of exacerbation was negative in each class, but magnitude varied.

**Keywords**

Classification, clustering, Dirichlet prior, joint model, longitudinal outcome, survival outcome, latent class model, medical monitoring

## 1 Introduction

Cystic fibrosis (CF) is a lethal genetic disorder that primarily affects the lungs. The clinical course of CF is marked by progressive loss of lung function and typically results in respiratory failure. Forced expiratory volume in 1 s (hereafter, $FEV_1$) is the most important clinical indicator in monitoring lung function decline in patients with CF. Patients during follow-up might experience acute respiratory events referred to as pulmonary exacerbations. It is, therefore, of clinical interest to characterize the association between the longitudinal outcome $FEV_1$ and time-to-first exacerbation. The motivation for our research comes from the US CF Foundation Patient Registry that consists of patients that were monitored from 2003 until 2015. In particular, we examined a subset of the Registry, which consists of 1016

[1]Department of Biostatistics, Erasmus MC, Rotterdam, The Netherlands
[2]Department of Hematology, Erasmus MC, Rotterdam, The Netherlands
[3]Division of Biostatistics & Epidemiology and Division of Pulmonary Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, USA
[4]Department of Pediatrics, University of Cincinnati, Cincinnati, USA

**Corresponding author:**
Eleni-Rosalina Andrinopoulou, Department of Biostatistics, Erasmus MC, PO Box 2040, 3000 CA Rotterdam, The Netherlands.
Email: e.andrinopoulou@erasmusmc.nl

patients. These patients were six years and older and were observed with a median number of follow-up visits equal to six (with a range of 1–93 visits). The average age at baseline is 15 years (with a range of 6–21).

Several authors have studied the evolution of lung function over time, as summarized in a recent review;[1] however, to our knowledge, little work has been done regarding the association of the lung function such as $FEV_1$ with time-to-event outcomes. In particular, joint modeling of longitudinal $FEV_1$ and survival outcomes in CF was introduced several years ago,[2] but has not been further used in CF epidemiology due to the computational burden of this approach. Furthermore, it is well recognized that different unobserved sub-groups of the biomarker $FEV_1$ exhibit different longitudinal profiles.[3] Patients can be categorized in several sub-groups (latent classes) with different trajectories. It is, therefore, of high clinical interest to measure the strength of association between $FEV_1$ with the risk of first exacerbation accounting for the latent trajectories.

The joint model of longitudinal and survival data constitutes a popular framework to analyze longitudinal and survival outcomes simultaneously.[4,5] In particular, two paradigms within this framework are the shared parameter joint models and the joint latent class models. The former paradigm links the longitudinal and the survival process via the random effects; however, it does not allow for latent classes.[6–11] The latter paradigm, which associates the two processes through latent classes, explicitly postulates the existence of sub-populations.[12–14] The main disadvantage of this approach is that there is no clear interpretation for the association of the two outcomes. In particular, it is not possible to obtain a parameter that quantifies the relationship between $FEV_1$ and time-to-first exacerbation. The use of latent classes in the shared parameter model has been previously proposed in the framework of joint models of longitudinal and survival data. In particular, Jacqmin-Gadda et al.[15] focused on the evaluation of the conditional independence assumption by proposing a score test; however, the relationship between the longitudinal and survival outcome per class is not further discussed.

In our clinical application, we are mainly interested in quantifying the association between the longitudinal outcome $FEV_1$ and the survival outcome time-to-first exacerbation using the shared parameter model. From the literature, it has been shown that sub-populations exist for the evolution of $FEV_1$.[3] The aim of the paper is twofold. Firstly, to model the relationship between $FEV_1$ and time-to-first exacerbation. For this purpose, we propose a Bayesian shared parameter joint model that integrates latent classes inherent in this heterogeneous population. This model will assess the strength of the association between the two outcomes while allowing for latent classes. Secondly, to address a problem that arises in latent class models, which is the selection of the optimal number of classes. Several approaches have been proposed in the literature both in frequentist and Bayesian frameworks, including among others the use of information criterion, Bayes factors, and reversible jump Markov chain Monte Carlo (MCMC). These approaches are computationally intensive and can require the fit of several models with different numbers of classes, which can be time-consuming. To overcome this problem, we will implement the method of Nasserinejad et al.[16] to our joint model. This method is a pragmatic extension of Rousseau and Mengersen[17] criterion that showed that when we overfit a mixture model by assuming more latent classes than present in the data, the superfluous latent classes will asymptotically become empty if the Dirichlet prior on the class proportions is sufficiently uninformative. Nasserinejad et al.[16] performed an extensive simulation study to further investigate this approach and used it as a criterion also in longitudinal studies for obtaining the optimal number of classes by simply excluding latent classes that are negligible in proportion.

## 2 Joint model estimation

### 2.1 Longitudinal submodel

To account for the fact that the population is heterogeneous and consists of $G$ possible unobserved sub-groups, we postulate a latent class mixed-effects model.[18–20] We let $\boldsymbol{y_i}$ denote the longitudinal response vector for the $i$th patient ($i = 1, \ldots, n$) obtained at different time points $t_{ij} > 0$ ($j = 1, \ldots, n_i$). In particular, we have

$$y_i(t|v_i = g) = \eta_{ig}(t) + \epsilon_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\beta}_g + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_{ig} + \epsilon_i(t) \quad (1)$$

where $v_i = g\,(g = 1, \ldots, G)$ presents the latent class indicator, $\boldsymbol{x}_i(t)$ denotes the design vector for the fixed effects regression coefficients $\boldsymbol{\beta}_g$, and $\boldsymbol{z}_i(t)$ is the design vector for the random effects $\boldsymbol{b}_{ig}$. Moreover, $\epsilon_i(t) \sim N(0, \sigma_y^2)$. For the corresponding random effects, we assume a multivariate normal distribution, namely

$$\boldsymbol{b}_{ig} \sim N(0, \Sigma_b)$$

where $N$ denotes the normal distribution and $\Sigma_b$ is the variance diagonal matrix of the random effects. An individual has a probability $\pi_{ig} = P(v_i = g)$ of belonging to latent class $g$. Using a multinomial distribution we obtain the class of each individual as

$$v_i \sim Multinomial(\pi_{ig})$$

According to the specification of the latent class mixed-effects submodel (1), both fixed and random effects are class-specific, whereas the measurement error $\epsilon_i(t)$ and the variance diagonal matrix of the random effects $\Sigma_b$ are not.

## 2.2 Survival submodel

We let $T_i^*$ denote the true failure time for the $i$th individual and $C_i$ the censoring time. Moreover, $T_i = \min(T_i^*, C_i)$ denotes the observed failure time and $\delta_i = \{0, 1\}$ is the event indicator where zero corresponds to censoring. We postulate a joint model for the relationship between the survival and the longitudinal outcome. Specifically, we have

$$h_i(t|v_i = g) = h_{0g}(t)\exp[\boldsymbol{\gamma}_g^\top \boldsymbol{w}_i + \alpha_g \eta_{ig}(t)] \tag{2}$$

where $\boldsymbol{w}_i$ is a vector of baseline covariates with a corresponding vector of regression coefficients $\boldsymbol{\gamma}_g$ and $h_{0g}(t)$ is the baseline hazard. Specifically, the B-splines baseline hazard function is assumed as $\log h_{0g}(t) = \gamma_{h_0 g,0} + \sum_{q=1}^{Q} \gamma_{h_0 g,q} B_q(t, \boldsymbol{v})$, where $B_q(t, \boldsymbol{v})$ denotes the $q$th basis function of a B-spline with knots $\nu_1, \ldots, \nu_Q$ and $\boldsymbol{\gamma}_{h_0 g}$ is the vector of spline coefficients. The knots are placed at the corresponding percentiles of the observed event times. Furthermore, $\alpha_g$ denotes the association parameter for the $g$th class. According to the specification of the survival submodel (2) the baseline covariates, the baseline hazard, and the association parameter are class-specific parameters. The proposed model goes beyond the standard joint model and joint latent class model where a single or no association parameter is assumed and provides a class-specific association. This is a more realistic assumption for the motivating data set since it is clinically expected that the risk of the first exacerbation will be higher when the rate of $FEV_1$ decline is faster. Accounting for these latent classes will lead to improved estimates of association arising from the joint model.

# 3 Bayesian estimation

We employ a Bayesian approach where inference is based on the posterior distribution of the parameters in the model. We use MCMC methods to estimate the parameters of the proposed model. The likelihood of the model is derived under the assumption that the longitudinal and survival processes are independent given the random effects and the latent classes. Moreover, the longitudinal responses of each subject are assumed independent given the random effects and the latent classes. The likelihood contribution for the $i$th patient is written as

$$p(\boldsymbol{y}_i, T_i, \delta_i | v_i = g, \boldsymbol{\theta}, \boldsymbol{b}_{ig}) = \sum_{g=1}^{G} \left\{ \pi_{ig} \prod_{j=1}^{n_i} [p(y_{ij}|v_i = g, \boldsymbol{\theta}_y, \boldsymbol{b}_{ig})] \times p[T_i, \delta_i | v_i = g, \eta_{ig}(T_i), \boldsymbol{\theta}_s, \boldsymbol{b}_{ig}] \right\}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_s^\top, \boldsymbol{\theta}_y^\top, \pi_{ig})^\top$ with $\boldsymbol{\theta}_y = (\boldsymbol{\beta}_g, \sigma_y, \Sigma_b)$ and $\boldsymbol{\theta}_s = (\boldsymbol{\gamma}_g, \alpha_g, \boldsymbol{\gamma}_{h_0 g})$.

The likelihood contribution of the longitudinal outcome takes the form

$$p(y_{ij}|v_i = g, \boldsymbol{\theta}_y, \boldsymbol{b}_{ig}) = (2\pi\sigma_y)^{-1/2} \times \exp\left[ -\frac{(y_{ij} - \boldsymbol{x_{ij}}^\top \boldsymbol{\beta}_g - \boldsymbol{z_{ij}}^\top \boldsymbol{b}_{ig})^2}{2\sigma_y^2} \right]$$

The likelihood contribution of the survival model is given by

$$p\{T_i, \delta_i | v_i = g, \eta_{ig}(T_i), \boldsymbol{\theta}_s, \boldsymbol{b}_{ig}\} = \exp[\gamma_{h_0 g,0} + \sum_{q=1}^{Q} \gamma_{h_0 g,q} B_q(T_i, \boldsymbol{v}) + \gamma_g^\top \boldsymbol{w}_i + \eta_{ig}(T_i)\boldsymbol{\alpha}_g]^{I(\delta_i = 1)}$$
$$\times \exp\{-\exp(\boldsymbol{\gamma}_g^\top \boldsymbol{w}_i) \int_0^{T_i} \exp[\gamma_{h_0 g,0} + \sum_{q=1}^{Q} \gamma_{h_0 g,q} B_q(s, \boldsymbol{v}) + \eta_{ig}(s)\boldsymbol{\alpha}_g, ]\mathrm{d}s\}$$

The posterior distribution is written as

$$p(\boldsymbol{\theta}, \boldsymbol{b}_g|\boldsymbol{y}, \boldsymbol{T}, \boldsymbol{\delta}) \times = \prod_{i=1}^{n} p(\boldsymbol{y}_i, T_i, \delta_i|v_i = g, \boldsymbol{\theta}, \boldsymbol{b}_{ig}) \times p(\boldsymbol{b}_{ig}|v_i = g, \boldsymbol{\theta}_y) p(\boldsymbol{\theta})$$

where

$$p(\boldsymbol{b}_{ig}|v_i = g, \boldsymbol{\theta}_y) = [2\pi \det(\Sigma_{bg})]^{-1/2} \times \exp(-\frac{\boldsymbol{b}_{ig}^{\top} \Sigma_{bg}^{-1} \boldsymbol{b}_{ig}}{2})$$

and $p(\boldsymbol{\theta})$ denotes the prior distributions.

A commonly used prior in mixture models for the class probability is a Dirichlet distribution. In particular

$$\pi_i \sim Dirichlet(\boldsymbol{a})$$

Small values of $\boldsymbol{a} = \{a_1 \dots a_G\}$ correspond to a less informative prior and a flat prior distribution is obtained when each $a_g$ is equal to 1. The selection of $\boldsymbol{a}$ is an important task and will be discussed in the next section. Standard priors can be assumed for the rest of the parameters. In particular, for the coefficients of the longitudinal fixed effects, the survival covariates, and the baseline hazard, normal priors can be taken. For the elements of the variance diagonal matrix of the random effects and the variance parameter of the longitudinal outcome, we can assume a gamma prior.

## 3.1 Selection of number of classes

An important task in latent class models is to identify the optimal number of classes. Several approaches have been previously proposed for choosing the optimal number of classes in both frequentist and Bayesian settings. Common examples are the Bayesian information criterion (BIC),[21] deviance information criterion (DIC),[22] and other Bayesian approaches such as Bayes factor and reversible jump MCMC algorithm.[23] A drawback of the approaches above is that they are computationally intensive and some require the fit of models assuming different numbers of classes, which might be time-consuming for complex models such as the joint models of longitudinal and survival outcomes.

Furthermore, it has been previously discussed in the literature that a problem that arises in the calculation of the DIC in mixture models is that the posterior mean of the parameters may not lead to good predictive estimates. The MCMC parameters suffer from label switching, making the DIC (which is based on averaging over MCMC draws) unstable. A more appropriate choice for the estimates would be the mode of the posterior distribution. A wide range of options for constructing an appropriate DIC, including also mixture models, has been proposed by Celeux et al.[22] However, these are not straightforward to apply with existing software.

An interesting alternative was proposed by Rousseau and Mengersen,[17] where they proved that in overfitted mixture models (with more latent classes than present in the data), the superfluous latent classes will asymptomatically become empty if the Dirichlet prior on the class proportion is sufficiently uninformative. Recently, Nasserinejad et al.[16] used this approach and proposed a latent class selection procedure for longitudinal models. An overfitted mixture model converged to the true mixture by assigning a small portion of individuals to empty classes, if the parameters of the Dirichlet prior $\boldsymbol{a}$ are smaller than $d/2$, where $d$ is the number of class-specific parameters (excluding the random effects). Furthermore, non-informative priors for the rest of the parameters are required. The steps are described as follows:

- First, a latent class model with a large enough number of latent classes is fitted.
- Then, the number of non-empty classes at each iteration is calculated as

$$g_{k,opt} = G - \sum_{g=1}^{G} I\left(\frac{n_{k,g}}{n} \leq \psi\right)$$

where $G$ is the total number of classes, $k$ represents the iteration, $n_{k,g}$ is the number of patients in class $g$ at iteration $k$, $n$ is the total number of patients, and $\psi$ is a predefined value.

- After obtaining the non-empty classes per iteration, the posterior mode of the non-empty classes is calculated.
- Finally, the model with the optimal number of classes, which are the non-empty classes, is refitted.

Advantages of this approach are that it is easy to implement even in such complex models, and it is not influenced by the label switching problem since we observe the non-empty classes at each iteration. The only time that we need to correct for label switching is when we fit the final model with the optimal number of classes. Furthermore, this approach requires us to fit the model only two times (namely one with the high number of classes and one with the optimal number of classes) instead of assuming all possible number of classes, therefore decreasing the computational burden. It has been shown through extensive simulations in the longitudinal setting that this method performs better than alternative model selection criteria such as BIC and DIC.[16]

## 4 Simulations

We performed a series of simulations to validate the proposed model, to examine the class selection method on the joint modeling framework, and to explore the appropriate threshold that indicates a class as empty.

### 4.1 Design

We assumed around 1000 patients with maximum number of repeated measurements equal to 10. For simplicity, in this section, we ignore the notation for the latent classes ($g$). To simulate the continuous longitudinal outcome, we used the following linear mixed-effects model per data set. In particular

$$y_i(t) = \eta_i(t) + \epsilon_i(t) = \beta_0 + \beta_1 \text{male}_i + \beta_2 t + b_{0i} + b_{1i} t + \epsilon_i(t)$$

where $\epsilon_i \sim N(0, \sigma_y^2)$ and $\boldsymbol{b}_i = (b_{0i}, b_{1i}) \sim N_2(\mathbf{0}, \Sigma_b)$. We adopted a linear effect of time for both the fixed and the random part, and corrected for a binary variable ($\text{male}_i$). Time $t$ was simulated from a uniform distribution between zero and 19.5. For the survival part, we assumed the following model

$$h_i(t) = h_0(t) \exp\{\gamma^\top \text{Age}_i + \alpha \eta_i(t)\}$$

The baseline risk was simulated from a Weibull distribution $h_0(t) = \xi t^{\xi - 1}$. For the simulation of the censoring times, an exponential censoring distribution was chosen so that the censoring rate was between 40% and 60%. Age was simulated from a normal distribution with mean 45 and standard deviation 15.7.

Under this setting, we simulated three different data sets that have different parameters for the fixed effects in the longitudinal submodel, the baseline covariates and baseline hazard in the survival submodel, the variance diagonal matrix of the random effects and the association parameter (more details are presented in Table 1). Figure 1 illustrates the evolution of the longitudinal outcome per gender from the simulation parameters for each one of the three data sets.

**Table 1.** Simulation parameters for the three data sets.

| | $\boldsymbol{\beta}$ | $\sigma_y$ | $diag\{\Sigma_b\}$ | $\xi$ | $\mu_c$ | $\gamma$ | $\alpha$ |
|---|---|---|---|---|---|---|---|
| Data set 1 | | | | | | | |
| | (Intercept) = 8.03 | 0.69 | 0.87 | 1.8 | 10 | (Intercept) = –4.85 | 0.38 |
| | Male = –5.86 | | 0.02 | | | Age = –0.02 | |
| | Time = –0.16 | | | | | | |
| Data set 2 | | | | | | | |
| | (Intercept) = –8.03 | 0.69 | 0.02 | 1.4 | 10 | (Intercept) = –4.85 | 0.08 |
| | Male = 12.20 | | 0.91 | | | Age = 0.09 | |
| | Time = 0.46 | | | | | | |
| Data set 3 | | | | | | | |
| | (Intercept) = 0.03 | 0.69 | 0.28 | 1.8 | 10 | (Intercept) = 2.85 | 0.58 |
| | Male = –1.96 | | 0.31 | | | Age = –0.12 | |
| | Time = –0.01 | | | | | | |

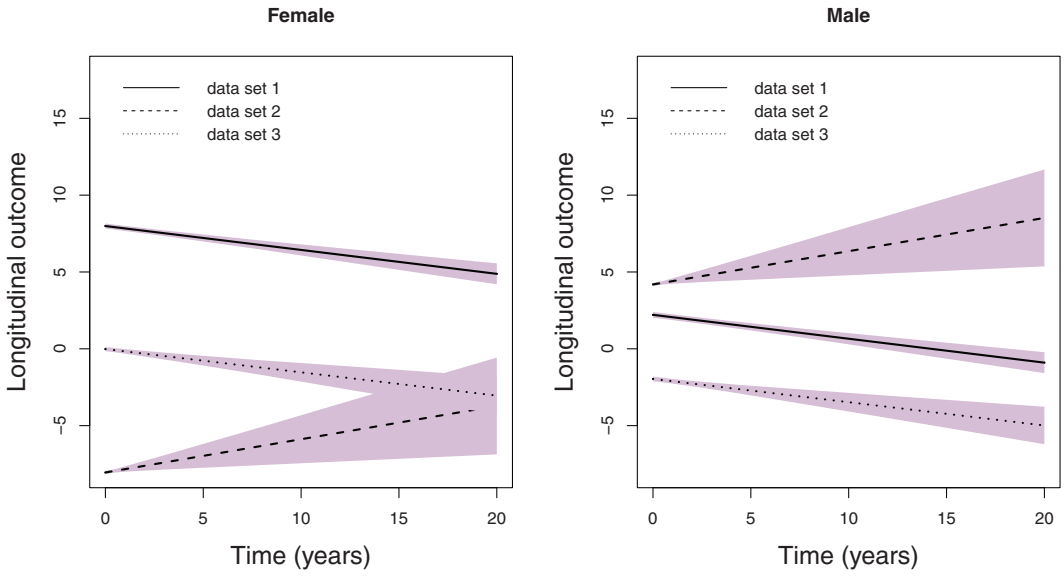


**Figure 1.** Evolution of the longitudinal outcome per gender from the simulation parameters for each one of the three data sets.

## 4.2 Analysis

In order to investigate the proposed methodology, we assumed the three following scenarios. For Scenario I, we combined all three data sets assuming $N_1 = 100$, $N_2 = 300$, and $N_3 = 500$ individuals. For Scenario II, we combined the two data sets with $N_1 = 300$ and $N_2 = 600$. Finally, for Scenario III, we used one data set with $N_1 = 900$. The separate data sets represent the sub-populations. We first used Scenarios II and III and fitted the proposed joint model (2) assuming two and one class, respectively, to evaluate the model. We, then, assumed a variety number of classes to investigate the proposed class selection method. In the fitted models, all parameters were class-specific except for the measurement error in the mixed-effects submodel and the variance diagonal matrix of the random effects. These include the fixed effects from the longitudinal submodel (three parameters), the baseline covariates from the survival model (two parameters), the baseline hazard (eight parameters) from the survival submodel, and the association parameter (one parameter). We assumed $a_g = 6.9$, which is smaller than the total number of the class-specific parameters divided by two. The priors that we used are:

- $\boldsymbol{\beta}_g \sim N(0, 1000)$,
- $\boldsymbol{\gamma}_g \sim N(0, 1000)$,
- $\boldsymbol{\gamma}_{h_0 g,q} \sim N(0, 1000)$,
- $\alpha_g \sim N(0, 100)$,
- $\sigma_y^2 \sim GA^{-1}(0.01, 0.01)$
- $\text{diag}\{\boldsymbol{\Sigma}_{bg}\} \sim GA^{-1}(0.01, 0.01)$,

where $GA^{-1}$ denotes the inverse gamma distribution. For the variance of the association parameter, no large value was required to ensure that we have a non-informative prior since we simulated this parameter to be smaller than 0.6. We ran the MCMC using a single chain with 50,000 iterations, 25,000 burn-in, and 10 thinning. We performed 150 simulations per scenario.

We compared our proposed class selection methodology with the DIC criterion. Several adaptations have been proposed for mixture models by Celeux et al.[22] The parameters are not always identifiable, and the posterior mean of the parameters can be a poor estimator; therefore, in our simulation study, we assumed $DIC_3$. A frequently used package, when the focus is on the association between a longitudinal and a survival outcome while taking into account that different sub-populations exist, is the `lcmm` in R developed by Proust-Lima et al.[24] We used the function `Jointlcmm()` and assumed the BIC criterion to obtain the optimal number of classes and compare it with the proposed approach. The same specifications, as described for the data simulation, were assumed for fitting the models. A difference can be found in the `Jointlcmm()` function, where the longitudinal and the survival outcomes are only connected via the latent classes, and a cubic M-splines baseline risk function was assumed. We present an overview of the simulation study in Table 2.

**Table 2.** Simulation study overview.

| | |
|---|---|
| Evaluating the proposed model | |
| Simulate | Fit |
| 1 data set | 1 class |
| 2 data sets | 2 classes |
| Evaluating the proposed class selection method (DIC)[a] | |
| Simulate | Fit |
| 3 data sets | 2, 3, 4 classes |
| 2 data sets | 1, 2, 3 classes |
| 1 data set | 1, 2 classes |
| Evaluating the proposed class selection method (BIC) | |
| Simulate | Fit |
| 3 data sets | 1–6 class |
| 2 data sets | 1–6 class |
| 1 data set | 1–6 classes |

[a]Note that for the DIC criterion, we did not assume all classes (1–6) since it is computationally intensive to run so many complex models.

## 4.3 Results

The results from the different scenarios for the validation of the proposed model are presented in Table 1 in the Supplementary Material. We obtain that the true parameters are close to the model parameters. The results from the different scenarios for the investigation of the proposed class selection approach are illustrated in Table 3. In particular, we present for different $\psi$ the percentage of the true number of classes and the mode of the number of classes.

For Scenario I, we obtain the highest percentage when assuming $\psi$ to be between 10 and 15%. In particular, assuming that the $\psi$ is equal to 15% we obtain 72% of the time the correct number of classes and a mode equal to the correct number of classes (three). On the other hand, using the DIC and BIC criterion, we obtain 7% and 13% of the time the correct number of classes, respectively. Furthermore, both methods seem to underestimate the true number of classes (mode equal to one).

For Scenario II, we obtain the highest percentage when $\psi$ is between 8 and 15%. In particular, assuming that the $\psi$ is equal to 15% we obtain 49% of the time the correct number of classes and a mode equal to the correct number of classes (two). On the other hand, using the DIC and BIC criterion, we obtain 0% and 5% of the time the correct number of classes, respectively. Similar to Scenario I, both methods seem to underestimate the true number of classes (mode equal to one).

Finally, for Scenario III, the DIC and BIC criteria seem to perform better than the proposed approach, where it almost always selects the correct number of classes (one). This is not surprising since those criteria always underestimated the true number of classes in the previous scenarios. These results are also in line with previous work by Nasserinejad et al.[16] Using the proposed approach and assuming that the $\psi$ is equal to 15%, we obtain 30% of the time, the correct number of class and a mode equal to two.

# 5 Analysis of the CF data

In this section, we present the analysis of the motivating data set. Our primary focus is to investigate the association between $FEV_1$ and time-to-first exacerbation by taking into account that we have sub-groups with different evolution over time for $FEV_1$. The first step is to obtain the optimal number of classes that can explain the heterogeneity of the population. From the literature, it is known that two or three classes are observed for the evolution of $FEV_1$ outcome.[3] Therefore, for the selection process, we fitted a joint model assuming six classes. For the longitudinal outcome, we assumed a linear mixed-effects submodel including natural cubic splines for time (modeled as age, in years) with one internal knot at 15.84 years (corresponding to 50% of the observed follow-up times) in both the fixed and random effects parts. The DIC criterion and subject-specific plots (illustrating the observed and predicted values) were used to investigate the need for a non-linear evolution over time. Based on the DIC criterion, the best fit was obtained when using cubic splines with two knots. However, the observed versus predicted value plots did not show any drastic difference when assuming one knot; therefore, we decided to continue with the simplified model. Furthermore, we corrected for some baseline characteristics which were

**Table 3.** Simulation results: cut-off $\psi$, percentage of true number of classes, mode of the number of classes.

| $\psi$ (%) | True # of classes (%) | Mode of # of classes |
|---|---|---|
| Scenario I: | | |
| 150 simulations | | |
| 3 classes simulated | | |
| 1 | 0 | 6 |
| 2 | 3 | 6 |
| 5 | 12 | 5 |
| 8 | 39 | 4 |
| 10 | 54 | 3 |
| 12 | 66 | 3 |
| 15 | 72 | 3 |
| Scenario II: | | |
| 150 simulations | | |
| 2 classes simulated | | |
| 1 | 10 | 6 |
| 2 | 27 | 2 |
| 5 | 30 | 3 |
| 8 | 35 | 2 |
| 10 | 37 | 2 |
| 12 | 44 | 2 |
| 15 | 49 | 2 |
| Scenario III: | | |
| 150 simulations | | |
| 1 class simulated | | |
| 1 | 0 | 5 |
| 2 | 2 | 4 |
| 5 | 8 | 2 |
| 8 | 15 | 2 |
| 10 | 20 | 2 |
| 12 | 22 | 2 |
| 15 | 30 | 2 |

mainly based on clinical relevance and the literature. These variables, together with descriptive statistics, are presented in Table 4. In Figure 2, the $FEV_1$ evolutions of 25 randomly selected patients with more than one repeated measurements are presented.

Specifically, the model takes the form

$$\begin{aligned} y_i(t) = \eta_{ig}(t) + \epsilon_i(t) = {} & \boldsymbol{\beta}_{0g} + \sum_{\omega=1}^{2} \boldsymbol{\beta}_{\omega g} \mathrm{ns}(\mathrm{Age_i}, \omega) + \boldsymbol{\beta}_{3g} \mathrm{Female} + \boldsymbol{\beta}_{4g} \mathrm{F508_{Homozygous}} \\ & + \boldsymbol{\beta}_{5g} \mathrm{F508_{Heterozygous}} + \boldsymbol{\beta}_{6g} \mathrm{F508_{Homozygous}} + \boldsymbol{\beta}_{7g} \mathrm{F508_{Neither}} + \boldsymbol{\beta}_{8g} \mathrm{SESlow} + \boldsymbol{\beta}_{9g} \mathrm{MRSA} + \boldsymbol{\beta}_{10g} \mathrm{MSSA} \\ & + \boldsymbol{\beta}_{11g} \mathrm{Pa} + \boldsymbol{\beta}_{12g} \mathrm{aspergillus} + \boldsymbol{\beta}_{13g} \mathrm{PancEnzymes} + \boldsymbol{\beta}_{14g} \mathrm{numVisityr} + \sum_{\omega=1}^{2} \boldsymbol{b}_{i\omega g} \mathrm{ns}(\mathrm{Age_i}, \omega) + \epsilon_i(\mathrm{t}) \end{aligned}$$

To investigate the association between $FEV_1$ and time-to-first exacerbation, we postulated the proposed joint latent class model

$$h_i(t, \boldsymbol{\theta}_s) = h_{0g}(t) \exp[\boldsymbol{\gamma}_g \mathrm{Gender}_i + \alpha_g \eta_{ig}(t)]$$

For the baseline hazard, we assumed a quadratic B-splines basis with five internal knots placed at the corresponding percentiles of the observed event times ranging from 11.7 until 20.8 years.

In the Dirichlet distribution for the prior of the class probability, following the recommendation in Nasserinejad et al.,[16] we assumed $\boldsymbol{a}$ smaller than $d/2$ (where $d$ is the number of class-specific parameters).

**Table 4.** Descriptive statistics of the variables that were used in the model.

| | Percentage |
|---|---|
| Gender | |
| Males | 43 |
| Females | 57 |
| Number of F508del alleles (genotype) | |
| Homozygous | 53 |
| Heterozygous | 32 |
| Neither | 6 |
| Missing | 9 |
| SESlow (using state/federal or having no insurance is a marker of low socioeconomic status) | |
| Yes | 48 |
| No | 52 |
| MRSA (methicillin-resistant *Staphylococcus aureus*) | |
| Yes | 16 |
| No | 84 |
| MSSA (methicillin-sensitive *Staphylococcus aureus*) | |
| Yes | 21 |
| No | 79 |
| Pa (*Pseudomonas aeruginosa*) | |
| Yes | 46 |
| No | 54 |
| Aspergillus | |
| Yes | 28 |
| No | 72 |
| PancEnzymes (taking a pancreatic enzyme supplement, marks pancreatic insufficiency) | |
| Yes | 40 |
| No | 60 |
| | Mean (standard deviation) |
| Numvisityr (number of visits at the last follow-up within the prior year) | 5 (3) |

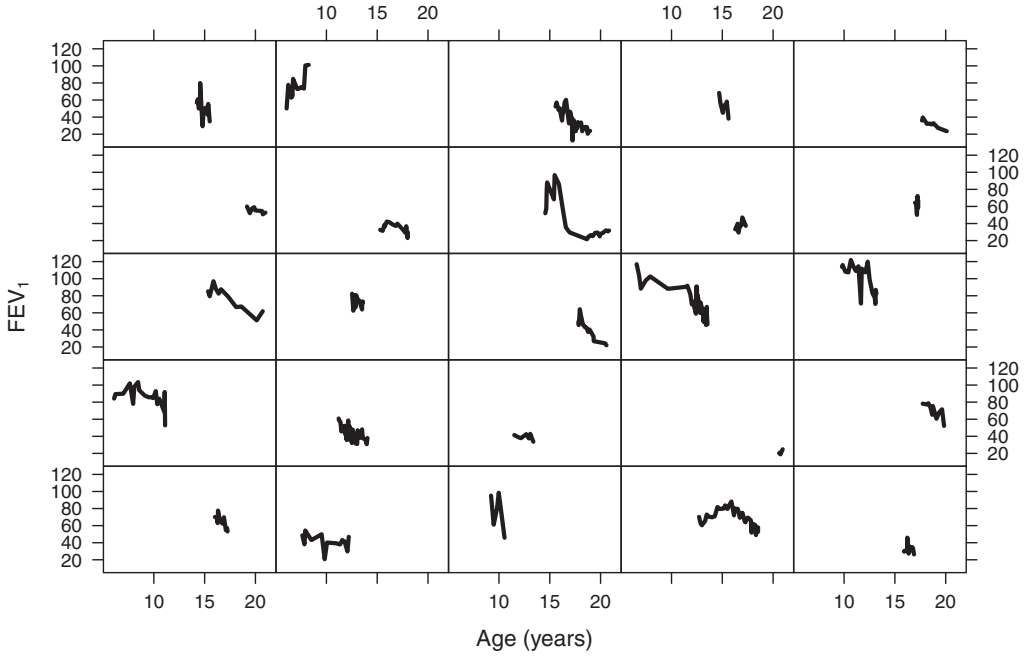


**Figure 2.** Individual $FEV_1$ evolutions of 25 randomly selected patients with more than two repeated measurements.

To ensure that we have the same scale for the coefficients of the covariates in order to easier select non-informative priors, we standardized the $FEV_1$ outcome and the continuous variables (age and numVisityr). Relatively uninformative priors were selected for the parameters in the model. These priors are as follows:

- $\boldsymbol{\beta}_g \sim N(0, 1000)$,
- $\boldsymbol{\gamma}_g \sim N(0, 1000)$,
- $\boldsymbol{\gamma}_{h_0 g,q} \sim N(0, 1000)$,
- $\alpha_g \sim N(0, 100)$,
- $\sigma_y^2 \sim GA^{-1}(0.01, 0.01)$
- $\text{diag}\{\boldsymbol{\Sigma}_{bg}\} \sim GA^{-1}(0.01, 0.01)$.

For the variance of the association parameter, no large value was required to ensure that we have a uninformative prior since with the standard joint model we obtained an association parameter smaller than 0.1. We ran the MCMC using a single chain with 500,000 iterations, 450,000 burn-in, and 100 thinning. The results indicate the presence of three or four classes, assuming that a class is empty if it contains 10 to 15% of the patients ($10\% \leq \psi \leq 15\%$). Since it is established in the literature that two or three classes are present in such populations, we decided to continue with three classes.[3,25]

We reran the model assuming three classes and the normal scale of the continuous covariate age, numVisitsyr and $FEV_1$ outcome. We ran the MCMC using 500,000 iterations, 450,000 burn-in, and 100 thinning. We, moreover, assumed two chains with different initial values to investigate the convergence towards local maximum. Density plots are presented in the Supplementary Material (Figures 1 to 5). We assumed the same priors as before except for $\boldsymbol{\beta}_g, \boldsymbol{\gamma}_g, \boldsymbol{\gamma}_{h_0 g,q}$, where we used a lower variance due to convergence problems, in particular we assumed $N$ (0, 100). We used a method proposed by Stephens[26] to handle the label switching problem. In particular, this method uses relabelling algorithms to perform a k-means type clustering of the MCMC samples. We should also note that when a lot of observations are available (like in our application), the label switching problem occurs less. Convergence was monitored by trace plots which are presented in the Supplementary Material (Figures 6 to 10). To investigate whether we have weak identifiability of the model parameters, we compared the prior and posterior distributions. The posterior was distinguishable from the prior indicating that our model is identifiable. The figures are presented in the Supplementary Material (Figures 11 to 15). Table 2 in the Supplementary Material shows the mean and standard deviation of age (at baseline), $FEV_1$ (at baseline), and number of visits (at last follow-up) per class, while Table 3 in the Supplementary Material shows the percentage of the categorical variables (at baseline) per class. In Figure 3, we illustrate the evolution of the longitudinal outcome in each class assuming patients who are females, are F508del homozygotes, without low SES, are not infections with MRSA, MSSA, or Aspergillus, do not use pancreatic enzyme, do not have *Pseudomonas aeruginosa* and had five visits within the prior year (which is the mean value of all observations). In general, we obtain a faster progression in class two. Patients in class three have a more stable evolution in the beginning of the follow-up and patients in class one seem to have an increased evolution in the beginning (this could be explained by the measurement error or possible transplantations). In addition, patients in class two start from a higher $FEV_1$ compared to patients in classes one and three. In Figure 4, we illustrate the evolution of the longitudinal outcome in each class assuming patients who are females, are F508del homozygotes, have low SES, are infections with MRSA, MSSA, Aspergillus, use pancreatic enzymes, have *Pseudomonas aeruginosa*, and had eight visits within the prior year. We obtain that patients in these classes start from a lower $FEV_1$ value compared to Figure 3. The mean and the credible interval of the MCMC samples of the association parameters per class are presented in Figure 5. In our proposed model, two of the three subgroups did not have a strong effect. In particular, we obtain a weak association between $FEV_1$ and time-to-first exacerbation for the first and third classes, while a strong negative association for class two. The association parameter $\alpha_g$ can be interpreted as the log hazard ratio of the time-to-first exacerbation for class $g$ when the $FEV_1$ value is increased by one unit, given that the gender is the same. In particular, for class two, a 10-unit decrease in the $FEV_1$ value (which is a clinical meaningful difference) results in a hazard ratio of 1.49.

We, furthermore, examined this association parameter while ignoring the latent sub-populations. In that case, the association parameter is smaller than the largest parameter from our proposed model. This indicates that the use of a common association parameter for all sub-populations would lead to an underestimated/overestimated parameter for each group. Since it is expected that patients with a constant lung-function trajectory are less likely to experience exacerbation compared to the patients with a steeper decline, it is not realistic to assume a common association between the $FEV_1$ evolution and time-to-first exacerbation for those group of patients.

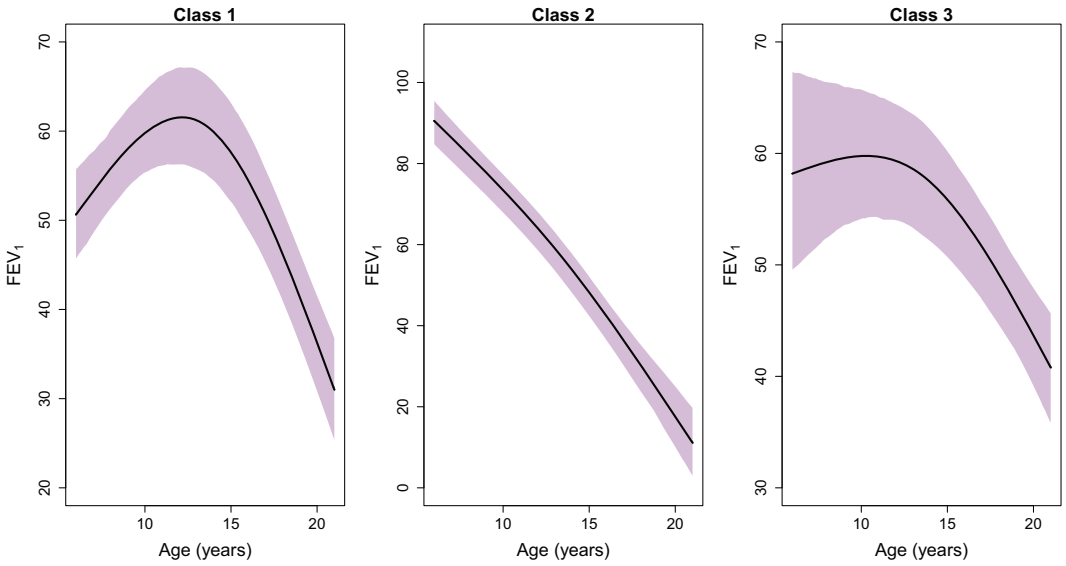


**Figure 3.** Evolution of the longitudinal outcome $FEV_1$ per class assuming patients who are females, are F508del homozygotes, without low SES, are not infections with MRSA, MSSA, or Aspergillus, do not use pancreatic enzyme, do not have *Pseudomonas aeruginosa* and had five visits within the prior year which is the mean value of all observations. The plots illustrate the posterior mean and credible interval for each class.

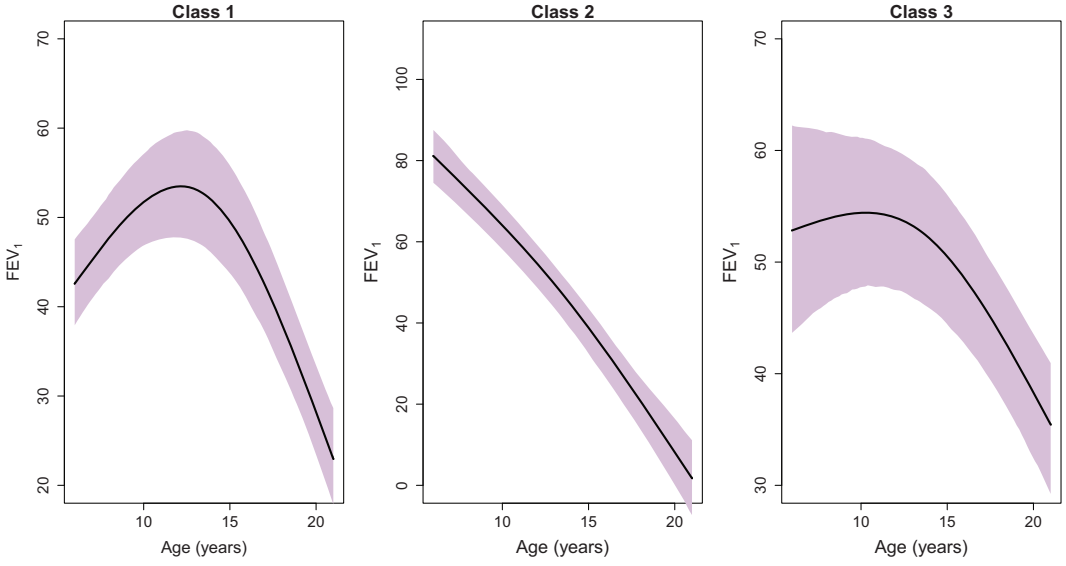


**Figure 4.** Evolution of the longitudinal outcome $FEV_1$ per class assuming patients who are females, are F508del homozygotes, have low SES, are infections with MRSA, MSSA, Aspergillus, use pancreatic enzymes, have *Pseudomonas aeruginosa* and had eight visits within the prior year. The plots illustrate the posterior mean and credible interval for each class.

## 6 Discussion

In this paper, we proposed a shared parameter joint model incorporating latent classes. Applying it to CF data, this model accounted for patient heterogeneity inherent in the progression of $FEV_1$. Compared to previously proposed joint latent class models,[13] we obtained the strength of the association between $FEV_1$ and time-to-first exacerbation per group of patients. Finally, we focused on the selection of the optimal number of classes and used an overfitted mixture model (high number of classes) to obtain the non-empty classes.

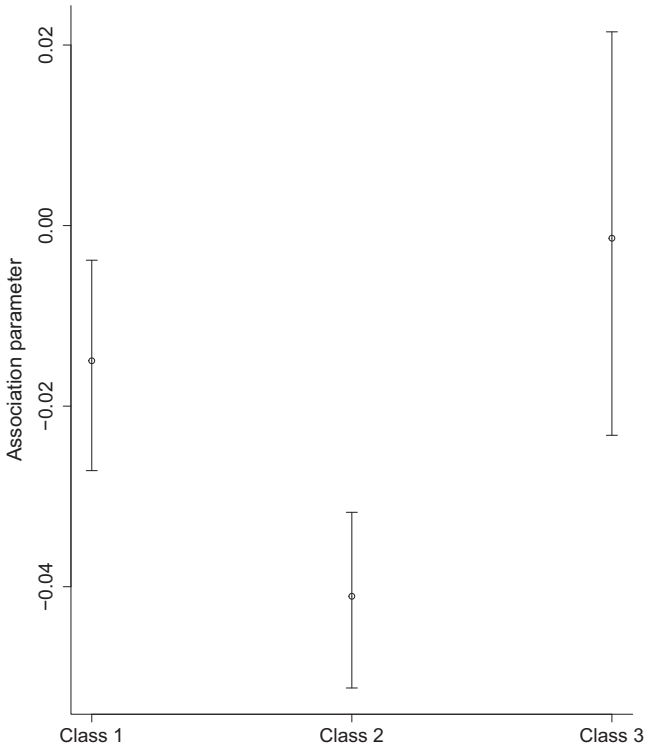


**Figure 5.** Mean and credible interval of the association parameter per class.

A limitation of this approach is that it requires an intensive computational effort. In particular, for the class selection, where a model with a high number of classes is required, the number of parameters increases drastically. This, in combination with the high number of observations in the CF application increases the computational time that is required. Considering the difficulty of this model, it is almost impossible to obtain the optimal number of classes with other Bayesian criterion. Implementing the proposed criterion is straightforward; however, due to the complexity of the model, it is computationally expensive to fit a model with a larger number of classes, e.g. 10. It was shown in the simulation analysis that the DIC and BIC criteria always underestimated the true number of parameters, and it, therefore, performed better when the true number of classes was one. Even though in that scenario, the proposed method did not work perfectly, it seems to be better than other criteria and easier to perform.

The main limitation and challenge of the proposed approach is the choice of the threshold that indicates whether a class is empty or not. A simulation study was performed to investigate whether the proposed class selection method would be appropriate for the shared parameter models. An interesting finding was that the threshold was higher compared to simple approaches, such as mixed models. In particular, in our shared parameter model with integrated latent classes, the threshold was between 10% and 15%. In more simple settings, such as a linear mixed model that was extensively investigated with simulations by Nasserinejad et al.,[16] this threshold was lower. Since this threshold highly depends on the complexity of the model, it is advisable to perform simulation studies to decide on a realistic cut-off when a different model is used. When no clear decision can be taken, our proposed criterion could be combined with other criteria to investigate the optimal number of classes in fewer models. Caution, however, is needed when using other criteria since we showed that the DIC and BIC performed worse in finding the correct number of classes. These results were in line with previous work by Nasserinejad et al.[16] The proposed approach led to the same number of classes as discussed in the CF literature on $FEV_1$ trajectory classification;[3,25] therefore, we did not investigate further the number of classes using other criteria. A further limitation of the manuscript is that we did not investigate which functional form describes best the association between $FEV_1$ and time-to-first exacerbation. In this work, we assumed the underlying values of $FEV_1$ to be related to time-to-first exacerbation as a first step to establishing an association between the two outcomes;[2,27,28] however, our future research focuses on investigating which functional form of $FEV_1$ is associated with the survival outcome exacerbations. Furthermore, in our model, recurrent exacerbations and death are not

taken into account. Although there is a large database available in the US Registry, we used only a subset in order to make it feasible to run the proposed model. This subset has particular characteristics, and it cannot be generalized to all patients in the Registry. Therefore, the presented results do not reflect the diversity of the whole database.

Possible extensions would be to include more covariates also in the survival submodel in order to take into account extra information regarding the patients. Furthermore, using the proposed model for obtaining future $FEV_1$ measurement and time-to-first exacerbation probabilities could lead to more efficient treatment prioritization and clinical management for patients with CF. In this paper, we used an overfitted mixture model and a non-informative Dirichlet prior for the class proportion in order to obtain the optimal number of classes. Overviews of mixture modeling, mixtures of Dirichlet processes, and how to identify the optimal number of classes using a Dirichlet prior can be found in Escobar and West and in Frühwirth-Schnatter and Malsiner-Walli.[29,30]

## Acknowledgements

The authors would like to thank the Cystic Fibrosis Foundation for the use of Cystic Fibrosis Foundation Patient Registry (CFFPR) data to conduct this study. Additionally, we would like to thank the patients, care providers, and clinic coordinators at Cystic Fibrosis centers throughout the United States for their contributions to the CFFPR. The authors would like to thank the two anonymous reviewers for critically reading the manuscript and suggesting substantial improvements.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by grants K25 HL125954 and R01 HL141286 from the National Heart, Lung and Blood Institute of the National Institutes of Health.

## ORCID iDs

Eleni-Rosalina Andrinopoulou https://orcid.org/0000-0002-5372-4163
Rhonda Szczesniak https://orcid.org/0000-0003-0705-715X

## Supplemental material

Supplemental material for this article is available online.

## References

1. Szczesniak R, Heltshe SL, Stanojevic S, et al. Use of fev1 in cystic fibrosis epidemiologic studies and clinical trials: a statistical perspective for the clinical researcher. *J Cystic Fibrosis* 2017; **16**: 318–326.
2. Schluchter MD, Konstan MW and Davis PB. Jointly modelling the relationship between survival and pulmonary function in cystic fibrosis patients. *Stat Med* 2002; **21**: 1271–1287.
3. Szczesniak RD, Li D, Su W, et al. Phenotypes of rapid cystic fibrosis lung disease progression during adolescence and young adulthood. *Am J Respir Crit Care Med* 2017; **196**: 471–478.
4. Tsiatis AA and Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Stat Sin* 2004; **14**: 809–834.
5. Hickey GL, Philipson P, Jorgensen A, et al. Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues. *BMC Med Res Methodol* 2016; **16**: 117.
6. Faucett CL and Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Stat Med* 1996; **15**: 1663–1685.
7. Wulfsohn MS and Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997; **53**: 330–339.
8. Brown ER and Ibrahim JG. Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* 2003; **59**: 686–693.
9. Rizopoulos D and Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat Med* 2011; **30**: 1366–1380.

10. Rizopoulos D. *Joint models for longitudinal and time-to-event data: with applications in R*. Boca Raton: Chapman and Hall/CRC Biostatistics Series, 2012.
11. Andrinopoulou ER, Rizopoulos D, Takkenberg JJ, et al. Joint modeling of two longitudinal outcomes and competing risk data. *Stat Med* 2014; **33**: 3167–3178.
12. Lin H, Turnbull BW, McCulloch CE, et al. Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *J Am Stat Assoc* 2002; **97**: 53–65.
13. Proust-Lima C, Séne M, Taylor JM, et al. Joint latent class models for longitudinal and time-to-event data: a review. *Stat Methods Med Res* 2014; **23**: 74–90.
14. Rouanet A, Joly P, Dartigues JF, et al. Joint latent class model for longitudinal data and interval-censored semi-competing events: application to dementia. *Biometrics* 2016; **72**: 1123–1135.
15. Jacqmin-Gadda H, ProustLima C, Taylor JM, et al. Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics* 2010; **66**: 11–19.
16. Nasserinejad K, van Rosmalen J, de Kort W, et al. Comparison of criteria for choosing the number of classes in Bayesian finite mixture models. *PLoS One* 2017; **12**: e0168838.
17. Rousseau J and Mengersen K. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J R Stat Soc Ser B (Stat Methodol)* 2011; **73**: 689–710.
18. Verbeke G and Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. *J Am Stat Assoc* 1996; **91**: 217–221.
19. Proust C and Jacqmin-Gadda H. Estimation of linear mixed models with a mixture of distribution for the random effects. *Comput Methods Programs Biomed* 2005; **78**: 165–173.
20. Proust-Lima C, Amieva H and Jacqmin-Gadda H. Analysis of multivariate mixed longitudinal data: a flexible latent process approach. *Br J Math Stat Psychol* 2013; **66**: 470–487.
21. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.
22. Celeux G, Forbes F, Robert CP, et al. Deviance information criteria for missing data models. *Bayesian Anal* 2006; **1**: 651–673.
23. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995; **82**: 711–732.
24. Proust-Lima C, Philipps V and Liquet B. Estimation of extended mixed models using latent classes and latent processes: the r package LCMM. *J Stat Softw* 2017; **78**: i02. Foundation for Open Access Statistics.
25. Moss A, Juarez-Colunga E, Nathoo F, et al. A comparison of change point models with application to longitudinal lung function measurements in children with cystic fibrosis. *Stat Med* 2016; **35**: 2058–2073.
26. Stephens M Dealing with label switching in mixture models. *J R Stat Soc Ser B (Stat Methodol)* 2000; **62**: 795–809.
27. Piccorelli AV and Schluchter M D Jointly modeling the relationship between longitudinal and survival data subject to left truncation with applications to cystic fibrosis. *Stat Med* 2015; **31**: 3931–3945.
28. Barrett J, Diggle P, Henderson R, et al Joint modelling of repeated measurements and time-to-event outcomes: flexible model specification and exact likelihood inference. *J R Stat Soc Ser B (Stat Methodol)* 2015; **77**: 131–148.
29. Frühwirth-Schnatter S and Malsiner-Walli G From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Adv Data Anal Class* 2019; **13**: 33–64.
30. Escobar MD and West M Bayesian density estimation and inference using mixtures. *J Am Stat Assoc* 1995; **90**: 577–588.